Data curation strategies to support responsible big social research and big social data reuse

Sara Mannheimer
Associate Professor - Data Librarian
Montana State University

IDCC 2022

Outline

Defining big social research

Key issues in big social research and big social data reuse

Data curation strategies

Big social research

- Data are large in scale

Big social research

- Data are large in scale
- Data collected from online sources—social media, blogs, forums

Big social research

- Data are large in scale
- Data collected from online sources—social media, blogs, forums
- Data collected using unobtrusive methods—APIs and web scraping

Big social research

- Data are large in scale
- Data collected from online sources—social media, blogs, forums
- Data collected using unobtrusive methods—APIs and web scraping
- Research conducted using computational social science methods—natural language processing, sentiment analysis, network analysis, AI, deep learning

This study

Literature review to identify key issues

Interviews about key issues

- Big social researchers
- Data curators

Key issues in big social research and big social data reuse

Key issues

- Context
- Data quality and trustworthiness
- Data comparability
- Informed consent
- Privacy and confidentiality
- Intellectual property and data ownership

Context

- Big social data are short pieces of text, images, or video, taken from a larger context of a person’s life

Context

- Big social data are short pieces of text, images, or video, taken from a larger context of a person’s life
- Out-of-context effect is compounded when data are amassed at a large scale.

Data quality and trustworthiness

- Missing data, sampling issues from APIs

Data quality and trustworthiness

- Missing data, sampling issues from APIs
- Bots and fake accounts

Data quality and trustworthiness

- Missing data, sampling issues from APIs
- Bots and fake accounts
- Representativeness of social media platforms

Data quality and trustworthiness

- Missing data, sampling issues from APIs
- Bots and fake accounts
- Representativeness of social media platforms
- Big social data are subject to loss over time—users can delete their accounts, links can become broken, and platforms can change

Data comparability

- Comparing and combining data can enhance context and quality of data

Data comparability

- Comparing and combining data can enhance context and quality of data
- Challenges:

Data comparability

- Comparing and combining data can enhance context and quality of data
- Challenges:
    - Matching participants across datasets

Data comparability

- Comparing and combining data can enhance context and quality of data
- Challenges:
  - Matching participants across datasets
  - Different data collection and sampling methods

Data comparability

- Comparing and combining data can enhance context and quality of data
- Challenges:
  - Matching participants across datasets
  - Different data collection and sampling methods
  - Different filetypes, metadata fields, metadata standards

Informed consent

- Scale of big social research makes it difficult to obtain informed consent from each user

Informed consent

- Scale of big social research makes it difficult to obtain informed consent from each user
- Social media terms of service may include consent clauses that cover big social research, but most users don’t read the ToS closely

Informed consent

- GDPR addresses consent, but not yet clear if it will in practice prevent “click and forget” consent systems

Informed consent

- GDPR addresses consent, but not yet clear if it will in practice prevent “click and forget” consent systems
- U.S. Dept of Health and Human Services has suggested using community focus groups and advisory boards to reduce harm

Informed consent

- GDPR addresses consent, but not yet clear if it will in practice prevent “click and forget” consent systems
- U.S. Dept of Health and Human Services has suggested using community focus groups and advisory boards to reduce harm
- None the big social researchers I spoke with had obtained informed consent from users

Privacy and confidentiality

- “Public” and “private” can be blurry online

Privacy and confidentiality

- “Public” and “private” can be blurry online.
- Big social researchers were concerned with privacy of research subjects, even through when data was collected from “public” platforms

IP and data ownership

- In the U.S., intellectual property on social media is a relatively gray area of the law

IP and data ownership

- In the U.S., intellectual property on social media is a relatively gray area of the law
- Social media companies view data as corporate assets

IP and data ownership

- In the U.S., intellectual property on social media is a relatively gray area of the law
- Social media companies view data as corporate assets
- Court cases invoking the Computer Fraud and Abuse Act to try to prevent web scraping

IP and data ownership

- In the U.S., intellectual property on social media is a relatively gray area of the law
- Social media companies view data as corporate assets
- Court cases invoking the Computer Fraud and Abuse Act to try to prevent web scraping
- Social media companies may limit data sharing (Tweet IDs only, for certain purposes only, etc)

Data curation to support responsible big social research

Data curation strategies

- Consultation throughout the research process
- Metadata and documentation
- Data repository services
- Advocacy for community standards

Consultation throughout the research process

- Encourage strategies for informed consent when possible—focus groups or automated consent requests

Consultation throughout the research process

- Encourage strategies for informed consent when possible—focus groups or automated consent requests
- Help researchers with rights management and navigating terms of service to support data sharing/reuse

Consultation throughout the research process

- Encourage strategies for informed consent when possible—focus groups or automated consent requests
- Help researchers with rights management and navigating terms of service to support data sharing/reuse
- Help researchers conduct risk-benefit analysis for big social research and big social data sharing.

Consultation throughout the research process

- Encourage strategies for informed consent when possible—focus groups or automated consent requests
- Help researchers with rights management and navigating terms of service to support data sharing/reuse
- Help researchers conduct risk-benefit analysis for big social research and big social data sharing.
  - e.g. balancing providing enough contextual information with protecting user privacy

Metadata and documentation

- Info about user communities

Metadata and documentation

- Info about user communities
- Research questions and research methods

Metadata and documentation

- Info about user communities
- Research questions and research methods
- Data collection, cleaning, analysis

Metadata and documentation

- Info about user communities
- Research questions and research methods
- Data collection, cleaning, analysis
- Potential errors, bias, missing data

Metadata and documentation

- Info about user communities
- Research questions and research methods
- Data collection, cleaning, analysis
- Potential errors, bias, missing data
- Related materials: software, code, related article DOI, related web links (using Perma CC)

Metadata and documentation

- Info about user communities
- Research questions and research methods
- Data collection, cleaning, analysis
- Potential errors, bias, missing data
- Related materials: software, code, related article DOI, related web links (using Perma CC)
- Metadata standards—DDI has been used, but no metadata standards specifically for big social data

Data repository services

- De-identification procedures

Data repository services

- De-identification procedures
- Restricted access

Data repository services

- De-identification (blinding usernames, paraphrasing, adjusting images)
- Restricted access
- Data enclaves

Data repository services

- De-identification (blinding usernames, paraphrasing, adjusting images)
- Restricted access
- Data enclaves
- Data use agreements

Data repository services

- De-identification (blinding usernames, paraphrasing, adjusting images)
- Restricted access
- Data enclaves
- Data use agreements
- Data licensing

Advocacy for community standards

- Big social researchers reported cobbling together strategies for responsible practice from many sources

Advocacy for community standards

- Big social researchers reported cobbling together strategies for responsible practice from many sources
  - Conducting on-the-fly risk-benefit analysis

Advocacy for community standards

- Big social researchers reported cobbling together strategies for responsible practice from many sources
  - Conducting on-the-fly risk-benefit analysis
  - Talking to colleagues and collaborators

Advocacy for community standards

- Big social researchers reported cobbling together strategies for responsible practice from many sources
  - Conducting on-the-fly risk-benefit analysis
  - Talking to colleagues and collaborators
  - Reading other studies

Advocacy for community standards

- Most big social researchers had not talked with data curators

Advocacy for community standards

- Most big social researchers had not talked with data curators
- It was rare for big social researchers to mention standardized ethical guidelines or community best practices

Advocacy for community standards

- Most big social researchers had not talked with data curators
- It was rare for big social researchers to mention standardized ethical guidelines or community best practices
  - (e.g. professional ethics guidelines and codes, AoIR Internet Research Ethics, Data Curation Network Primers)

Advocacy for community standards

- Most big social researchers had not talked with data curators
- It was rare for big social researchers to mention standardized ethical guidelines or community best practices
    - (e.g. professional ethics guidelines and codes, AoIR Internet Research Ethics, Data Curation Network Primers)
- Data curators can help advocate for and connect researchers with community standards

Conclusion

Data curators have tools, services, and knowledge that can help support responsible big social research and data sharing

Conclusion

Data curators have tools, services, and knowledge that can help support responsible big social research and data sharing

Tailor services to meet the needs of big social data

Conclusion

Data curators have tools, services, and knowledge that can help support responsible big social research and data sharing

Tailor services to meet the needs of big social data

Connect more with big social researchers to support responsible research and data sharing practices.

Thanks!

sara.mannheimer@montana.edu
saramannheimer.com